ZKG INTERNATIONAL

ISSN: 2366-1313

ENHANCING CREDIT CARD FRAUD DETECTION SYSTEMS WITH MACHINE LEARNING ON APACHE SPARK

#1KOMMIDI NARAYANA REDDY,
#2KALAKONDA VARSHINI,
#3SADULA SANKEERTH, *Assistant Professor,*
Department of Computer Science and Engineering,
SREE CHAITANYA INSTITUTE OF TECHNOLOGICAL SCIENCES, KARIMNAGAR, TS.

ABSTRACT: Credit cards, smartphone micropayments, and app cards are increasingly replacing cash as the primary method of payment. Because of the surge of online crime, banks utilize Fraud Detection Systems (FDS) to protect their consumers from fraudulent online banking. Using real-time information on users and funds, this technique effectively detects suspicious activity. Despite the fact that FDS has been demonstrated to reduce theft, the majority of the incidents it uncovers are False Positives, which necessitate costly investigations and leave consumers angry. The purpose of this research is to identify solutions to improve the current procedure. Researchers modified and merged the distributions of testing and training courses to evaluate what effect they had. Patterns and payment histories were mined as part of our research to help us identify unusual transactions. We also discussed the big picture of the data mining method utilized to uncover unusual credit card purchases. Using Python and Apache Spark, we were able to manage massive amounts of data rapidly and easily.

Keywords: credit card, Fraud detection, Outlier detection, GBT classifiesr

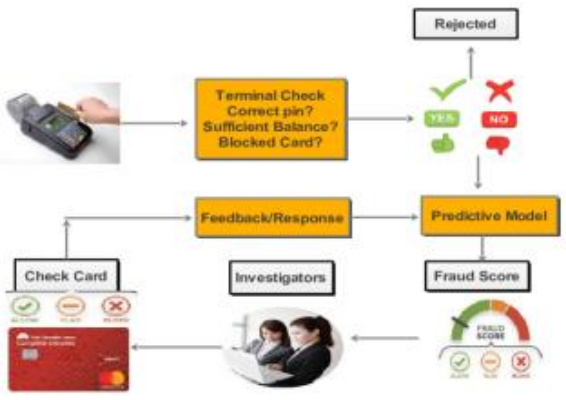
1. INTRODUCTION

Because there are so many point-of-sale (POS) machines, credit cards are widely utilized, both online and offline. People are going about their daily lives all around you. Con artists find it easy to con people because they are careless. Identity theft becomes popular when people give crooks access to their personal and private information. We need to facen in order to determine what type of transaction fraud has occurred.

There are numerous issues. It may be difficult to discern the difference between genuine and expired bargains.

Credit cards are a popular method of payment all around the world, but especially in North America. Credit cards with 0% interest for one month are extremely popular due to their use. Credit cards, like any other valuable instrument, can be abused. At the same time, credit card theft and extortion are on the rise. Every year, financial institutions (FIs) lose a lot of money and fall victim to complex fraud schemes. According to data from throughout the world, Visa and MasterCard lose more than $1 billion per year to frauds.

Credit card firms and charitable organizations are developing innovative methods to combat fraud. Cards are protected by magnetic strips, 3D logos, and card verification codes (CVC). Credit card firms are also developing an electronic smart card that will eventually replace plastic cards. People in the United States already believe that this move will fail since there are too many POS systems and card kinds available. The Financial Intelligence System (FIS) detects suspicious activity and prevents it from being investigated further by employing various computer approaches such as neural networks (NNs).



ISSN: 2366-1313

Figure 1: Account Charge Permission

Related work

The authority to charge an accountNagi and colleagues develop a data mining-based system for detecting financial fraud and security breaches. The works were published in 49 different periodicals in 2018. This article discusses four types of fraud and six distinct ways to employ data mining.Sanjeev and his colleagues examine and categorize fraud types, transaction quantities, and money lost by country using real credit card transaction data. Following that, a box plot is used to demonstrate how dispersed the data is.That is why Michael et al. developed a signature-based study technique and a fraud detection model to assess how effective real-time fraud detection is.Real-time recognition necessitates both speed and precise accuracy.

PROBLEM STATEMENT

You must also plan extensively for the data extraction process. That is, Quah and his associates. It was divided into four layers: the "identity" layer, the "inspection" layer, the "core" layer (which examined risk scores and behavior), and the "SOM algorithm" review layer. This allows for immediate identification.Even if a purchase appears to be fraudulent, it may not be. Many individuals have heard the term "false positive," which refers to a case that was initially classified as suspicious but turned out to be true. Because the user must always grant permission for any transaction that isn't usual, the consumer is more likely to be dissatisfied. It could be highly costly to investigate a large number of misleading findings.

Motivation

Every day, many genuine cases (FPs) must be examined, which requires a significant amount of time and work. If there were fewer FP reviews, fraud analysts could devote more time and attention to genuine fraud cases. This strategy can assist banks and other financial institutions in reducing their losses.

2.SYSTEM ANALYSIS

Proposed System

The primary purpose of this research is to devise a method for categorizing lists of actual and false items into groupings of more likely real items. If this is implemented, it will be much easy to investigate numerous potentially fraudulent transactions. Financial organizations can save a lot of money by reducing the number of hours they spend on unnecessary study. Many fraud situations that are currently undiscovered would be discovered if the present FDS threshold was reduced. This allows people to intervene earlier to prevent losses. Outlier detection and the GBT Classifier, two of the most prominent Machine Learning approaches for discovering patterns and categorizing them, can help to overcome these issues. The results demonstrate that the adopted strategy has the potential to significantly improve the current system.

Advantages

- The answer solves the problem of incorrect forecasts.
- Artificial intelligence is utilized in modern fraud detection algorithms to help weed out false positives.

Functional Requirements

The solution is simple and quick. The features that enable the object to function are what make it helpful. The functional needs of a project are the specifics of what it must do.

Non-Functional Requirements

Non-functional considerations such as how it appears, how safe it is, how easy it is to use, how reliable it is, and how fast it is may be important to the client.

Performance Requirements

- Your speed metrics indicate how quickly the software reacts to your input. There should be no more than three seconds between app launches.
- The entire process of validating the information should take no more than five seconds.
- The results can be sent in no more than five seconds.

Design Constraints- PyCharm will be the Windows IDE used to construct and release the

ISSN: 2366-1313

Python program.

Standards Compliance - Variable names should be used consistently throughout a program. Users must be satisfied with what they see and feel. We require an easy-to-understand and use virtual user experience.

Reliability- The item must not be damaged during the process.

Availability- You can always make use of the application.

Security- If a service retains personally identifiable information about its users, it must do so in a secure and safe manner.

Maintainability- The program's administrators should be skilled in data management.

Portability- It should work with any version of Windows.

Hardware Requirements:

You'll need a Core i3 processor and at least 4 GB of RAM.

You have access to up to 20 GB of PC storage.

The computer screen, keyboard, and mouse are examples of secondary devices.

Software Requirements:

This is true for both Windows 8 and Windows 10. Python 3.6 and the PyCharm IDE (integrated programming environment) are utilized.

APIs include the Numpy, Pandas, PySpark, and Matplotlib packages.

3.SYSTEM DESIGN

System Architecture:

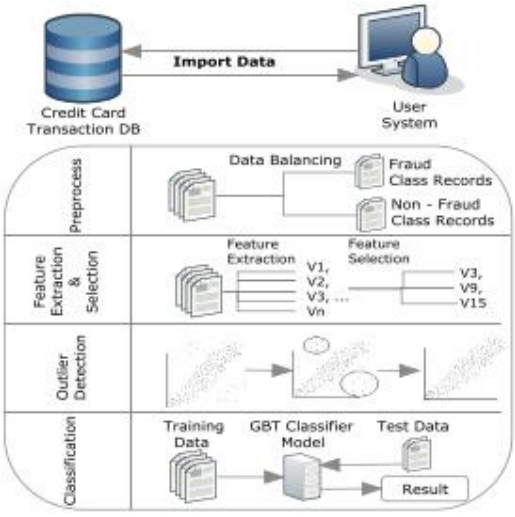


Figure 2: Making Things Work

The diagram below depicts how the CCFDS operates. Customers can use this method to keep credit card transaction databases up to date. It is critical to determine why credit card theft has recently increased.

Data collection: Obtain the financial information required by the raw dataset.

Data balancing: When you have adequate information, divide it into two categories: class-0 signifies no fraud and class-1 means fraud.

Feature extraction and selection: Class 1 data revealed 492 fraudulent trades. This project is currently working on versions 1-2 and 28.

Outlier detection: It determines the space between each pair of data in order to prepare for grouping. "Outliers" are numbers that appear in data sets that were not used for training.

Classification: Given the asymmetry of the data, many models tend to prefer the most prevalent class. PySpark, like SQL, can examine massive volumes of already organized or mostly organized data. The GBT Classifier uses it to sort and group received data lines. An Illustration of Abuse and Misuse

Use Case diagram

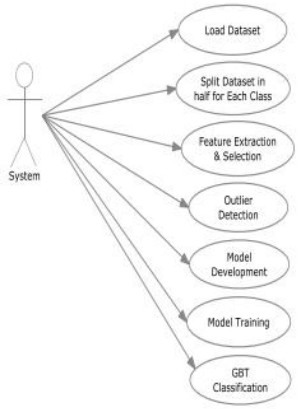


Figure 3: Use Case Diagram

DFD (Data Flow Diagram)

The DFD facilitates communication between the machine and its user. The entire project life cycle is depicted below. To find bargains, you must do three things: We'll begin by investigating data fusion. The next step is to categorize the information.

ISSN: 2366-1313

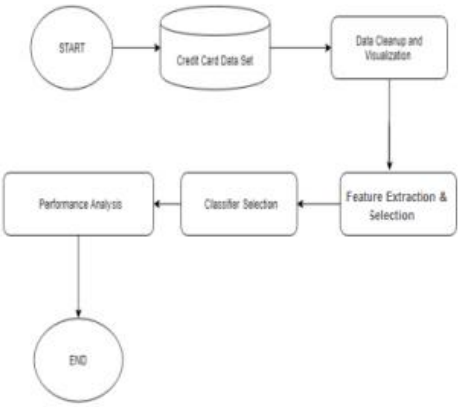


Figure 4: DFD in First Grade

4.SYSTEM IMPLEMENTATION

The suggested method is divided into sections. Here are some of the most important aspects of the method described.

Module Names

- Data Collection
- Data Balancing
- Feature Extraction
- Outlier Detection
- Classification

Module Description

1. **Data Collection:** In the last two days, there have been 2,84,807 credit card purchases. This data set contains 492 false records, which is more than twice as many as true ones (0.17%). Keep silent about the improved versions 1, 2, 3, and V28 available. We can detect the symptoms of fraud (a value of 0) and the lack of fraud (a value of 1) in the very last column.
2. **Data Balancing:** Unbalanced classes occur frequently in ML-based sorting. This occurs when one or more classes include far too many instances. This is because most ML systems are designed to be more precise and less prone to errors. In this scenario, there is no distinction between the groupings or the number of them. It's difficult to set a benchmark because only 492 of the 2,84,077 transactions have been flagged as potentially fraudulent. Using Python's pandas package, the number of genuine transactions must be reduced till it equals the number of forgeries.
3. **Feature Extraction:** While the heat map method can help you determine what distinguishes one class from another, it also makes the recognition system less sensitive. A heat map depicts the major and minor values of a grid as a dispersion of brilliant dots. In this scenario, we group the matrix's rows and columns together. The traits are rated, and the most useful for training the model are chosen.
4. **Outlier Detection**: The distance between data points is considered in both cluster analysis and outlier detection. It is, however, utilized to highlight facts and opinions that contradict the remainder of the material. An outlier is a number that deviates significantly from the rest of the data. We should aim to make as few mistakes as possible when training the model. This is accomplished using the Python package Numpy.
5. **Classification:** Classification has numerous applications. This statement can be applied to any situation in which a decision or prediction is made based on the information provided. It works for the above categories because they have specific characteristics or characteristics. The purpose is to generate an idea that can be utilized to determine which of the present classes an observation belongs in. Because it requires a sample with known true classes, the process of creating a classification system is also known as pattern recognition, discrimination, and supervised learning. Data is sorted using GBT Classifier, and data is sent using PySpark. PySpark, like SQL, can examine massive volumes of already organized or mostly organized data. The GBT Classifier uses it to sort and group received data lines.

5.TEST CASES

ISSN: 2366-1313

TEST CASE NUMBER	TEST	OUTPUT	RESULT
1	Open project in PyCharm	Project loads successfully	Pass
2	Run the main file	Program executes	Pass
3	Data loading function is executed	Data is loaded	Pass
4	Data balancing function is called	Data is split in equal proportion as fraud and non-fraud data	Pass
5	Feature Extraction is performed	Features are extracted from the data	Pass
6	Feature selection is performed	Heatmap is generated and moist significant features are selected	Pass
7	Outlier detection function is called	Outlier data is detected and discarded	Pass
8	Classification function is called	Results are classified as fraud and non-fraud data	Pass
9	Accuracy is derived from the function	Accuracy is measured and displayed	Pass

6. RESULT AND ANALYSIS

The initial step is to install Python, the integrated development environment, and the 3.6 pycham files**.**

Make a separate folder for every of your credit cards**.**

The final step is to make PyCharm's IDE available for download**.**

Fourth, the company begins to use the credit card information**.**

The fifth stage is to write code that can be used to detect credit card fraud**.**

Step 6: Review the code and correct any errors you find**.**

The seventh step is to execute the written code**.**

Examine to check if there are just 492 incorrect entries out of a total of 284807.

There have been no cases of fraud. 28.4315 out of 79 points

Which of the following is correct: 77 False discoveries occur 4.6% of the time, which is a significant issue.

The correct answer is 0.92941765758824.

The odds are one in 0.9518072289158826.

Based on which criteria: 92.9411764

Fig5: Sort the commodities by type.

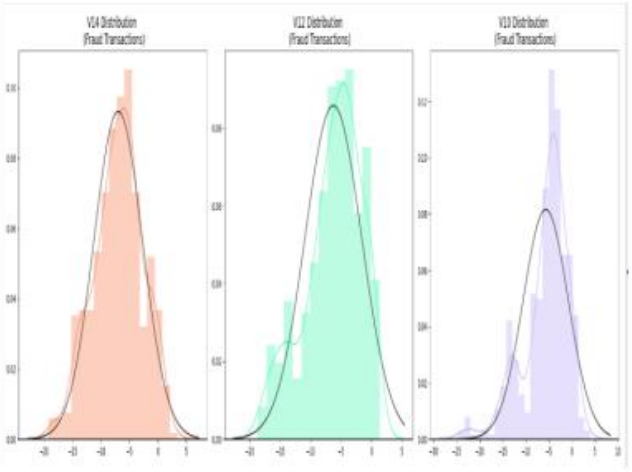


Fig6: Different methods for detecting financial wrongdoing

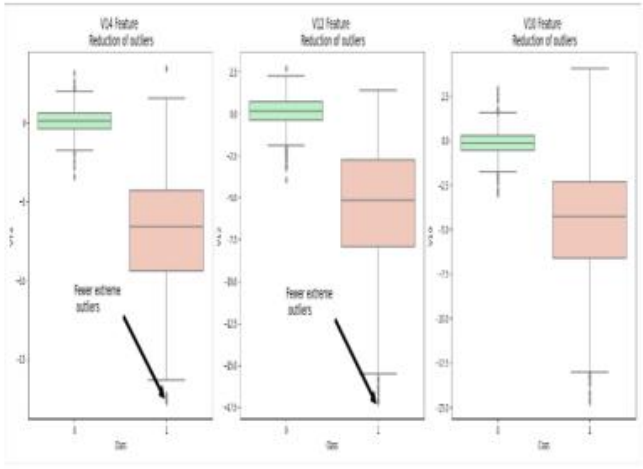


fig 7: Getting rid of everything that could do you harm

7. CONCLUSION

Because there are more convenient ways to pay and internet money transfers, the likelihood of a fraudulent payment has increased. Credit card

ISSN: 2366-1313

fraud includes the fraudulent use of a credit card, forgery, and theft of personal information associated with a lost or stolen card. Credit cards are most commonly stolen by phishing, pharming, and keeping private card information out in the open. As a result, the government established a "e-financial fraud prevention service." Keyboard filters, public licensing, or a second password will not prevent financial theft. The system immediately examines the user's ID and payment information, notifies both the user and the financial institution, and suspends the transaction if any issues are discovered. That is why more research is needed to refine the algorithm and develop a method to accurately detect suspicious financial activities. The purpose of this study was to see if it was possible to detect false activity using machine learning and electronic payment data analysis. The results suggest that the procedures utilized on the data were effective, and the categorization was proper.

Deep learning apps that employ neural networks may improve accuracy in the future. More datasets could be utilized to ensure that the methods presented operate.

REFERENCES

[1] Donald V. Macdougall, Richard G. Mosley, Garioch J. l. Saunders; Credit card crime in Canada: Investigation - Prosecution; The Canadian Association of Crown Counsel; page 1-56; January 1985.

[2] Isabelle Sender; Detecting and combating fraud; Chain Store Age; New York; Vol. 74; Issue 7; Page 162; July 1998.

[3] Elford Dean, Raj Thomas, Lorry; Visa security center; Personal meetings; January 7 and February 11,1999.

[4] Gyusoo Kim and Seulgi Lee, “2014 Payment Research”, Bank of Korea, Vol. 2015, No. 1, Jan. 2015.

[5] EWT Nagi, Yong Hu, HY Wong, Yijun Chen, Xin Sun, “The Application of Data Mining Techniques in Financial Fraud Detection: A Classification Framework and an Academic Review of Literature,” Decision Support Systems, Vol. 50, No. 3, Feb. 2011.

[6] Jha, Sanjeev, J. Christopher Westland, “A Descriptive Study of Credit Card Fraud Pattern,” Global Business Review, Vol. 14, No. 3, pp. 373-384, 2015.

[7] Edge, Michael Edward, Pedro R. Falcone Sampaio, “A Survey of Signature based Methods for Financial Fraud Detection,” Computers & Security, Vol. 28 No. 6, pp. 381- 394. 2009.

[8] Aihua Shen, Rencheng Tong, Yaochen Deng, "Application of Classification Models on Credit Card Fraud Detection," Service Systems and Service Management of the 2007 IEEE International Conference, pp. 1-4, Jun.2007.

[9] Chengwei Liu, Yixiang Chan, Syed Hasnain Alam Kazmi, Hao Fu, “Financial Fraud Detection Model: Based on Random Forest,” International Journal of Economics and Finance, Vol. 7, No. 7, pp. 178-188, 2015.

[10] Ganesh Kumar.Nune and P.Vasanth Sena, “Novel Artificial Neural Networks and Logistic Approach for Detecting Credit Card Deceit,” International Journal of Computer Science and Network Security, Vol. 15, No. 9, Sep. 2015.